

An Efficient Approach to Segmentation of Cell-Nuclei: Application in Breast Cancer Tissue Blocks

PS Umesh Adiga*, Adam Idica, Rodrigo Gonzalez, Carlos Ortiz de Solorzano
Lawrence Berkeley National Laboratory
1, Cyclotron Road, Berkeley, CA94720
upadiga@lbl.gov

Abstract: This paper presents a software system developed to study large tissue blocks of Ductal Carcinoma *In Situ* (DCIS) and to quantify genetic variation in them. Given the size of the image data set and the characteristics of the images, direct application of any single global algorithm to segment the cell-nuclei and quantify genetic heterogeneity is a difficult exercise. The software system we propose is based on an integrated approach where several image analysis algorithms are integrated under one robust methodology to extract the maximum amount of information from the tissue blocks. The study also shows how 2-D image analysis of a large number of nuclei can be effective in observing genetic heterogeneity in cancerous tissue blocks.

Keywords: Tissue, Nuclei, Segmentation, Genetic Heterogeneity.

I. Introduction

This paper presents an integrated approach for automatic segmentation and analysis of genetically aberrant cell-nuclei in large tissue blocks of breast cancer tissue. The study of these large data sets is useful in understanding the genetic heterogeneity of malignant tumors. One such case is the analysis of Ductal Carcinoma In Situ (DCIS) of the breast, a malignant, pre-invasive form of breast cancer. DCIS is characterized by an abnormal proliferation of transformed cancer cells within the boundaries of the existing epithelial ductal structures. Holland *et al.* [1], performed histo-pathological gross and microscopic analyses of mastectomy specimens from 214 women to show that 30% of the cases had an extensive intra-ductal component (EIC+), as defined by the presence of intra-ductal tumor foci, more than 2cms away from the primary tumor. The presence of multiple foci is also reported by Lagios *et al* [2] in DCIS cases without tumor invasion, suggesting that intraductal

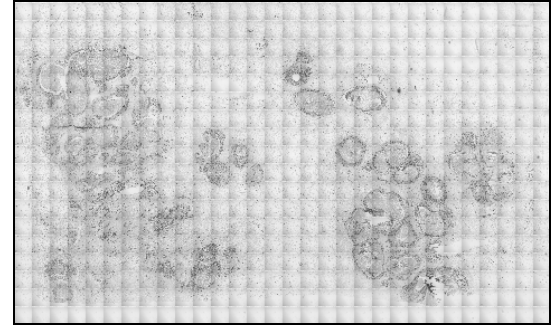
spreading can occur before stromal invasion. Determining whether these secondary foci are in fact morphologically and/or genetically related to the primary tumor requires reconstructing the tumors in 3D, determining the continuity of the ductal structures and then comparing the genetic make-up of the cells of the foci and the primary tumor. We hypothesize the existence of a preferential intraductal invasion and therefore a common origin between the primary tumor and the secondary foci. This would give credence to the hypothesis that DCIS may exist extensively along the ducts at a significant distance from the tumor and that it may progress to invasive disease later on. Verification of such a hypothesis involves analyzing a large number of tissue specimens with significant tumor and ductal areas as shown in Figure 1(a). After reducing the background pattern by subtracting a prototype of an empty snapshot as shown in Figure 1(b), important regions of interest such as ducts and tumors are marked either automatically. The corresponding regions in the fluorescent stained sections are reacquired at a higher magnification. A small part of the fluorescent tissue image acquired at higher magnification is shown in Figure 1(c).

Automatic evaluation of histological images exists on several levels. Complete automation, semi-automation and task-specific automation systems. In the present case, each tissue block consists of hundreds of data sets of an average size of $5K \times 5K$ pixels, and an average nuclear density of 4000 nuclei/image. This makes complete automation not only desirable but also essential. Different techniques to segment cell nuclei in tissue images have been reported since the 1970s. A review of most of the early techniques is provided by Preston and Bartels [3]. However, in spite of the development of multiple techniques for image segmentation, there is no reliable and

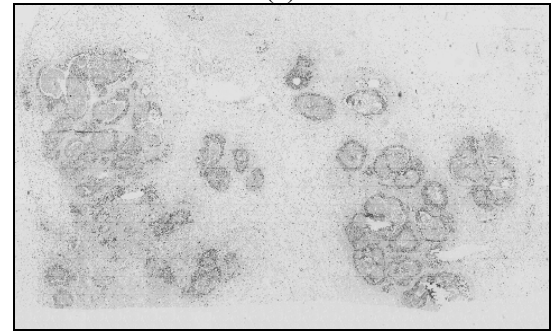
global technique that can be successfully applied for segmentation of histopathological image analysis. Gill *et al* [4] in their review of image analysis and morphometry in breast cancer have concluded that the absence of any reliable nuclear segmentation technique is the main cause of the lack of automation in tissue image analysis. Nedzeved *et al* [5] and Schupp *et al* [6] proposed morphological filters for cell-nuclei segmentation in tissue images. Schupps *et al* [6] have constructed an integrated approach based on several morphological filters. They report a good success in segmentation but the rate of false positives is as high as 35% in some cases, with an average of 11%. Barba and Gill [7] have proposed a parametric model fitting algorithm for cell segmentation. In it, they assume that objects are convex in shape and hence that a shape model can be used for segmentation. However, in images where not only different cell types are present, but also normal and transformed cells can be found, the assumption of a standard shape for cell nuclei might not provide good segmentation. Yang and Parvin [8] have used a quadratic shape model whose parameters can be relaxed for the segmentation of non-standard shapes. Zimmer *et al* [9] used active contour models for the segmentation of cell-nuclei. Considering the number of nuclei to be segmented, initializing the active contour models would pose a significant problem. In addition active models are prone to noise and other artifacts present in the tissue image. The thesis on quantitative evaluation of 3-D images of tissue sections by Umesh Adiga [10] provides an in-depth study of the practical problems involved in large-scale segmentation of multi-dimensional and multi-spectral tissue images. In the present work we have tried to solve the problems that rendered many of the earlier methods only a project specific. We have done this by extensive pre-processing to standardize the images for segmentation and then by recovering the shape of each nucleus by a shrink-wrap process.

A robust approach that automatically extracts regions of interest and segment at least 75% of cell nuclei with less than 10% false positives would thus be essential for quantitative analysis. We have not come across any publication on segmentation of

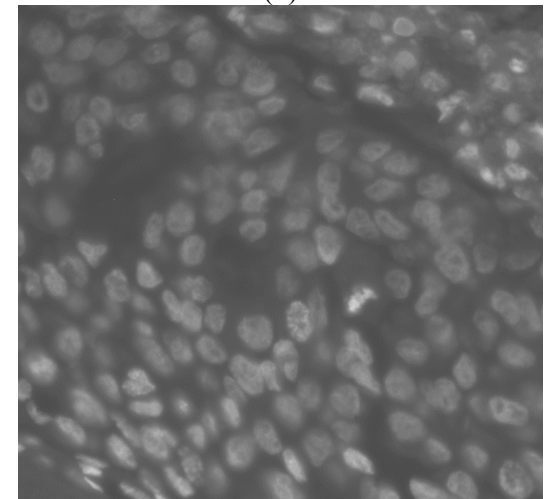
large tissue data that integrates several filters and can be used on different images without substantially changing algorithm parameters, *i.e.* fully automatic.



(a)



(b)



(c)

Figure 1: (a) Complete tissue section image (b) After background pattern removal (c) A small part of the region of interest in corresponding fluorescent tissue section.

The rest of the paper is presented as follows. Section II describes the methods designed and implemented to automate quantification of the desired features. Section III describes how genetic heterogeneity in DCIS breast tissue blocks is quantified and section IV provides experimental results and discussion.

II Methods

Large tissue blocks are physically sectioned into $5\mu m$ thick sections. Alternative sections are stained with Hematoxylin and Eosin (H&E) and the rest of the thin sections are stained using a nuclear counter-stain and fluorescence in situ hybridization targeting gene *Her2* and centromere specific chromosomal locations. Low magnification (2.5X) images of the H&E and fluorescent sections are first acquired and approximately registered with each other to understand the spatial connection of the ducts, tumors, etc. Regions of interest (ROI) are semi-automatically extracted in each H&E stained tissue image and are virtually mapped onto its neighboring fluorescent-stained tissue image. The multi-spectral images of the regions of interest in the fluorescent stained tissue specimens are reacquired at a higher magnification. The blue color channel contains the tissue structure with clearly recognizable cell nuclei. Red and green color channels contain fluorescent signals corresponding to the centromere (green) and a specific area of chromosome 17 (red), known to host the *Her2* gene, which is commonly amplified in DCIS tumors. The tissue image in the blue color channel is segmented and the result of segmentation is mapped onto red and green channels and the genetic variation of *Her2* and the centromere of chromosome 17 within the cell nuclei are recorded.

The image acquisition and analysis software system is built in two parts. The first part facilitates image acquisition and semi-automatic marking of regions of interest such as tumors, milk ducts, etc in a low-resolution image. It also facilitates automatic registration of the slices to form a complete spatial structure of ducts and tumors. The mapping of regions of interest to fluorescent-stained tissue images and reacquiring regions of interest at higher resolution is done by this image acquisition system. The details of the first part have been reported earlier [11]. The second part is the analysis software system that performs pre-processing, segmentation nuclei, recognition of properly segmented and completely represented nuclei and recording of genetic variations within the nuclei.

In histo-pathological images, the features that facilitate segmentation of individual cell nuclei are image gradient peaks representing nucleus boundary and the deep boundary/surface concavities where more than one cell-nucleus touch or appear to overlap. A segmentation process preceded by techniques aimed at enhancing these two features would thus contribute towards a more robust and generally acceptable segmentation. A hierarchical, multi-stage segmentation process is essential in such cases. The control flow diagram of the integrated approach is shown in Figure 2

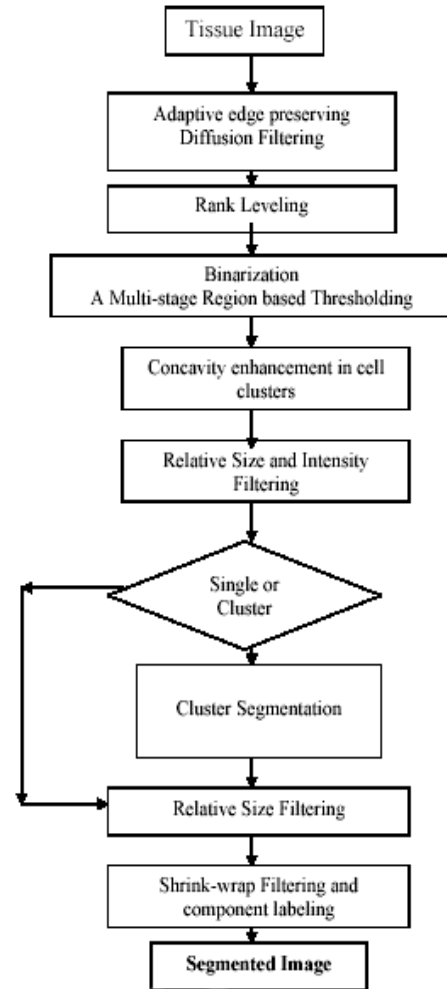


Figure 2: Flow diagram of a generalized segmentation process for histo-pathological images

Figure 2: Flow diagram of a generalized segmentation process for histo-pathological images

A. Pre-processing

The electronic instrumentation noise, uneven illumination due to camera vignetting, improper staining of the tissue, etc., are the main causes of noise and artifacts. We have

implemented a reaction-diffusion filter to smooth the image to reduce random noise while simultaneously enhancing the boundary features of the objects [12]. Consider the image function

$$U = I(x, y) \quad \text{----- (1).}$$

The partial differential equation for the reaction-diffusion filter can be written as

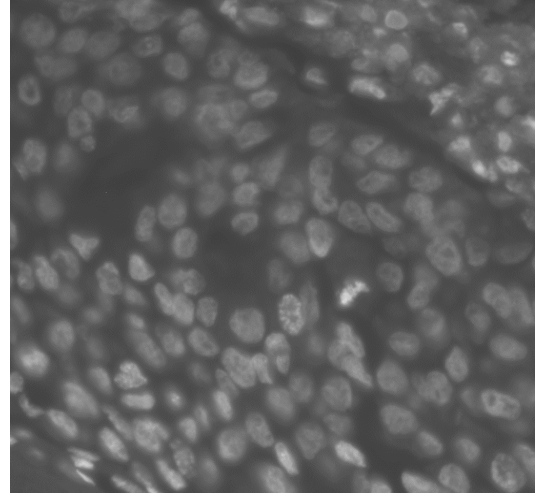
$$\frac{\partial U}{\partial t} = (\cos \beta) \cdot \nabla h \cdot \nabla U + (\sin \beta) \cdot h \cdot \nabla^2 U \quad \text{----- (2)}$$

where $h = 1 / \left(1 + \left(\frac{\partial I}{\partial x} \right)^2 + \left(\frac{\partial I}{\partial y} \right)^2 \right)$ is the

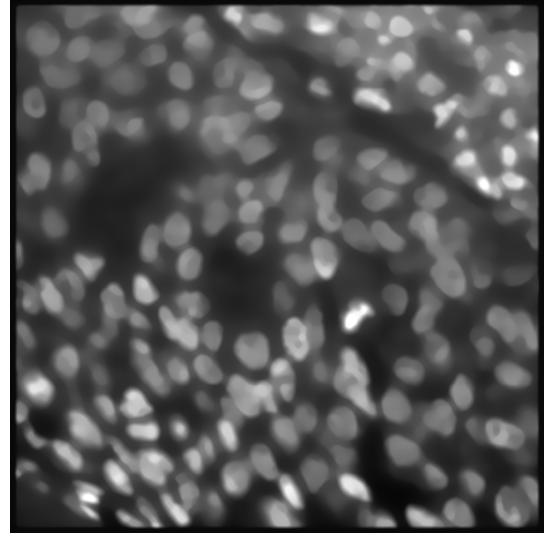
gradient function and β is a parameter that determines the relative contribution of reaction and diffusion terms and varies from 0° to 90° . This formulation provides a minimum diffusion at the edges and extensive diffusion elsewhere [13]. The first term, $\{(\cos \beta) \cdot \nabla h \cdot \nabla U\}$ is a reaction term responsible for edge-enhancement while the second term $\{(\sin \beta) \cdot h \cdot \nabla^2 U\}$ is a diffusion term responsible for smoothing. $\beta = 0^\circ$ is a pure reaction term, $\beta = 45^\circ$ is a nonlinear diffusion flow, $\beta = \arctan 2 \approx 63.4^\circ$ is the Beltrami flow, and $\beta = 90^\circ$ is the pure diffusion. Figure 3 shows the result of adaptive smoothing using action diffusion filter mentioned above.

Due to partial absorption of the light or camera vignetting, improper alignment of the light-source and the camera, variation in tissue specimen thickness, etc., the images are darker at the corners than at the center. We have used a rank-leveling approach to remove uneven background illumination. Rank leveling is an adaptive, multi-step morphological filtering process. In the first step, a background image is constructed by diffusing the image to an extent that the objects are melted into the background. This can be accomplished by using grey-scale erosion or by repetitive smoothing. The resulting image is an approximate representation of the background and the overall brightness variation. The background image is then subtracted from the original image, and the grey values below zero are

clipped to zero. The grey-levels of the image are then stretched to improve the contrast. This process tends to normalize the data sets by standardizing histogram shape and average brightness.



(a)



(b)

Figure 2: (a) A small area of the tissue section (b) After adaptive smoothing.

The next stage of the processing is to separate foreground pixels from the background. This is accomplished by automatic region-based thresholding.

B. Automatic Thresholding

Separation of the objects and the object clusters from the background can be achieved by appropriate thresholding of the grey-scale image. Thresholding is a process of conversion from a grey-scale image to a bi-level image. A bi-level image should contain all the information concerning the number, position and structure of the objects that are present in the grey-level image while containing much less other information. The

problem is to select a proper threshold that accomplishes the above task. We have observed that selection of a single threshold for an entire image is not possible in most cases due to overlapping of the grey-levels of the objects and background from different regions of the image. Thus the first issue that must be addressed, if regional thresholds are to be used, is to determine distinct regions and the corresponding threshold levels. We have implemented a multi-step, adaptive region selection method to solve this problem.

In the first step, the smoothed image is amplitude thresholded at a global mean intensity value given as,

$$\tau_{global} = k \cdot \mu \quad \text{-----} (3)$$

where k is a tuning constant and μ is the global mean intensity of the image. The tuning parameter can be set experimentally for a batch of images acquired under similar microscope and illumination settings. In our experiment we have used $k = 1$ as a default value. Many global threshold selection techniques can be used to replace this initial step [14]. It is our experience that average image intensity is normally a good initial guess for a preliminary stage of thresholding. The experimental results of the standard global thresholding methods such as exhaustive-search [15], entropy-based thresholding [16], histogram-shape based thresholding [17] were not significantly different. The aim of the first step is to divide the image into different regions so that the further thresholding of the regions at a unique threshold value is possible. Hence, an approximate thresholding at an average intensity value is considered as a first step instead of using more complex methods.

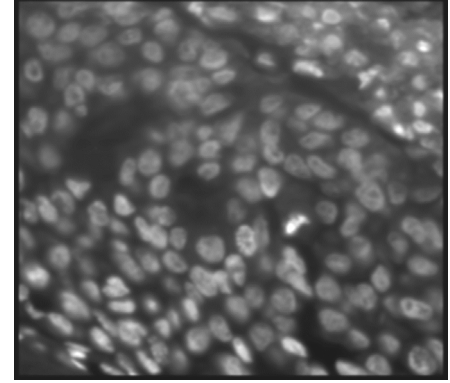
In the second step, all the connected components in the foreground of the initial thresholded image are identified by component labeling. Let N be the number of unique components in the initial thresholded image.

In the third step, the mean grey-levels μ_i for each connected component $i = 1, 2, 3, \dots, N$ are calculated from the grey-level image.

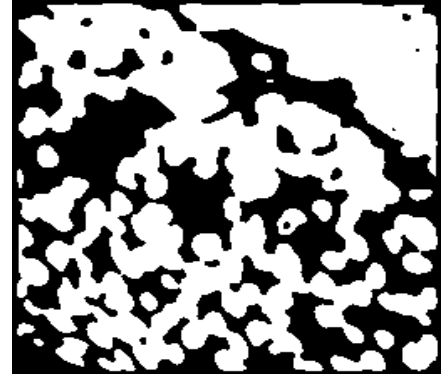
In the fourth step, each connected component i is thresholded at a unique threshold value

$$\tau_i = c \cdot \mu_i \quad \text{-----} (4).$$

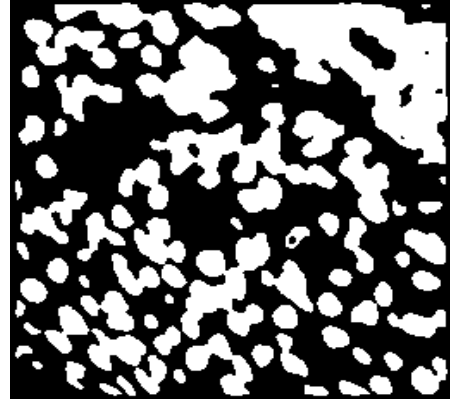
The tuning factor c is experimentally set (default $c = 0.5$).



(a)



(b)



(c)

Figure 4: Result of multi stage region thresholding (a) Original image (b) Result of first stage of thresholding (c) Result of second stage of thresholding

In this way we not only divide the image into unique regions but we also calculate a unique threshold value for each and every region. The method is simple but very effective in distinguishing different regions in the image and setting proper thresholds based on the local region brightness. Figure 4 shows the result of multi-level thresholding on a small part of the tissue image.

To reduce the number of holes within the nuclei as well as spurious offshoots at the object surface, we applied a grey scale morphological operations such as ‘closing’ and ‘opening’ over amplitude thresholded and component labeled image using a structuring element with an approximately circular effective kernel. The circular shape of the structuring element reduces noisy concavities and offshoots in the object surface. This structural smoothing of the object boundary reduces the fragmentation of the nucleus during subsequent steps of segmentation of the cell nuclei clusters.

C. Segmentation of nuclei clusters

Segmentation of the cluster of nuclei is the most difficult aspect of histological image analysis. Absence of strong boundary features where nuclei appear to touch or overlap on one another is a main reason for difficulty in delineating the nucleus boundary. Our approach to segmentation is thus focused on enhancing features that facilitate accurate marking of the nucleus boundary leading to segmentation. The boundary concavities that exist where cell-nuclei touch one another, is one such feature that should be enhanced. This is done as follows.

In the first step a local gradient magnitude image is obtained as a first derivative of the image intensity value. A general gradient map is shown in Figure 5(a).

In the second step, the gradient image is subject to a boundary sharpening filter. We have modified a directional smoothing filter to smooth selectively along the boundary while suppressing the non-boundary gradient values [15]. At every pixel location in the gradient map, the local 3×3 neighborhood is divided into four directions and the average gradient magnitude in each direction is calculated. The direction of the boundary at a pixel location is that direction in which the average gradient magnitude is maximum. A semi-Olympic average of the gradient intensity in the boundary direction is then used as the value of the pixel in the smoothed gradient map. An Olympic averaging discards extreme values while averaging. In semi-Olympic filter we have discarded only the maximum value. This process not only facilitate suppression of

noisy gradient peaks but also implicitly facilitate sharpening of the gradient peaks at the boundary by suppressing the gradient magnitude of the pixels located away from the boundary. Figure 5(b) shows the result of sharpening the gradient magnitude image by the above described process.

Next step is to synthesize the nuclei boundary from the sharpened gradient image. All the pixels in this smoothed gradient image with gradient value smaller than the average gradient magnitude of the image is suppressed to zero value. From the resulting image, the primary, secondary and tertiary local gradient peaks are retained for further processing. The primary gradient peak is the pixel with a maximum gradient magnitude in a 3×3 neighborhood. The secondary peak and tertiary peaks are the second and third maximum gradient magnitude values in a 3×3 neighborhood. The gradient peaks image is smoothed and the skeleton of a smoothed gradient peak image can be expected to provide an approximate, though discontinuous, boundary of the cell-nuclei. The whole process of gradient peak selection, smoothing, etc., is necessary to avoid the effect of noisy gradient peaks. If they are left unchecked, they result in fragmentation of the objects in clusters during final stages of segmentation. Figure 5(c) shows the skeleton of the gradient map superimposed on the original grey-level image.

In the fourth step, all those pixels in the thresholded image that correspond to the skeleton of the gradient map are converted into background pixels. This creates concavity, where cell nuclei touch or appear to overlap and hence vastly improves the segmentation result. This process, even if does not completely separate touching cells, enhances the features that indicate the overlap of cell nuclei.

At this stage, some of the cell nuclei might touch one another by a fairly thin connection and need a simple pinching-off by a few iterations of erosion or boundary peeling followed by conditional dilation or thickening to be separated. Ultimate erosion and conditional dilation for complete segmentation of the clusters as proposed by Russ [18] has many practical disadvantages.

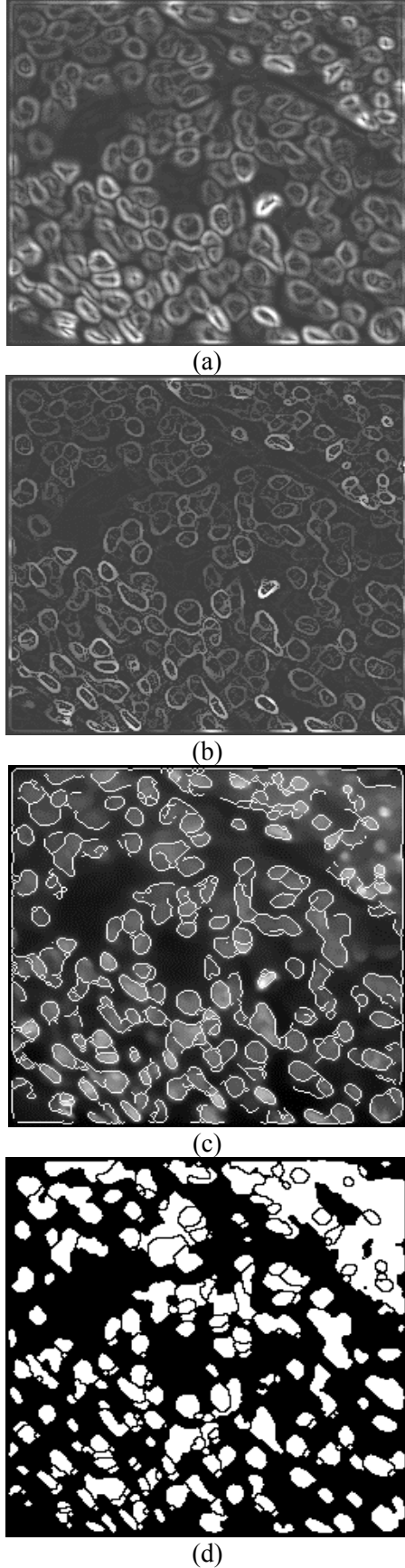


Figure 5: (a) Gradient magnitude image (b) After directional enhancement (c) Boundary synthesized from the gradient image is (d) Binary image after concavity enhancement.

The assumptions of similar size and shape for the objects in the cluster, distinguishing noisy signatures from the proper signature of the nucleus, stopping erosion of the signatures selectively to avoid loss of information, keeping track of the number of iterations of erosion each signature has gone through to become a unique signature, etc., are some of the disadvantages of using iterative erosion and dilation technique.

When the approximate size and the average intensity values of the nuclei in the image are known, this information can be provided as *a priori* information to the filtering process for selecting individual isolated cell-nuclei and flagging cluster of nuclei for further processing. Otherwise, a data-driven process to filter-out the artifacts is necessary. We have used relative size and relative intensity filters to eliminate artifacts and flag-off object clusters for further processing and to remove artifacts. The relative size of the object r_v , defined as the ratio of the size of that object to the average size of objects in the image, is used to eliminate artifacts and separate isolated objects from clusters. All individual objects are first ordered based on their sizes i.e. number of pixels within the objects. The average size is calculated by an α -trimmed filter which excludes α number of extreme size elements in the size ordered list of objects for calculation of the average object size in the image. We have used $\alpha = 25\%$ thus cutting of 25% of the objects on both extremes of the size-ordered and average intensity ordered list of the cell-nuclei respectively. If the size of the object i is V_i , then the relative size of the object i is given by

$$r_{v_i} = \frac{V_i}{\frac{1}{(N - 2\alpha)} \cdot \sum_{k=\alpha}^{N-\alpha} V_k}$$

where N is the number of isolated objects present in the image and α is the cut-off threshold of the α -trimmed filter. The relative intensity of the object r_{I_i} is defined

as the ratio of the average intensity of the object pixels to the average intensity of

foreground pixels in the image. If the average intensity of the object i is \bar{I}_i , then

$$r_{I_i} = \frac{\bar{I}_i}{\frac{1}{(N-2\alpha)} \cdot \sum_{k=\alpha}^{N-\alpha} (\bar{I}_k)}$$

The relative object size and relative object intensity of each individual object in the image is calculated. All those objects with relative mean object intensity less than 0.3 are considered as artifacts and eliminated. Objects whose relative size is above 1.1 and below 0.7 are flagged off for second stage of segmentation. The numerical thresholds mentioned here are set based on experiments on considerably large number of data and are automatically updated. Let, I_{seg} be the image containing segmented cell nuclei and $I_{cluster}$ the image containing the clusters of cell nuclei that are yet to be segmented. A region growing technique that can be argued as a generalized watershed algorithm is initiated for cluster segmentation [19].

A distance map of $I_{cluster}$ is generated by using a distance-transform as explained by Borgefors [20]. This distance map provides a reconstructed grey-scale image of $I_{cluster}$ with local intensity peaks representing the centers of the objects in the cluster and the grey-level decreases uniformly from the center of the objects towards its boundary. Region growing algorithm on this distance map is described in a few steps as follows.

Let $dist(.)$ represent the distance value of pixels in the distance map, d_{max} be the maximum distance in the distance map, d_{next} be the next maximum distance level and d_{min} be the minimum distance value in the distance map.

Step 1: The pixels having maximum distance d_{max} in the distance map are considered as regional markers. The markers are labeled by a sequential component labeling algorithm. To start with, a regional marker may consist of a single pixel or a group of connected pixels.

Step 2: Pixels having a distance value (d_{next}) and located in the immediate

neighborhood of the regional markers are merged with the corresponding regional marker.

This step can also be viewed as growing the regional markers into their neighborhood pixels that have a distance value d_{next} by tagging them with the corresponding label of the marker.

The pixels or the groups of connected pixels in the distance map with distance d_{next} and not having a labeled regional marker in their immediate neighborhood are considered as new markers and are given new labels. At the end of this step all the pixels with distance value d_{next} have a unique label attached to them showing they belong to regional markers.

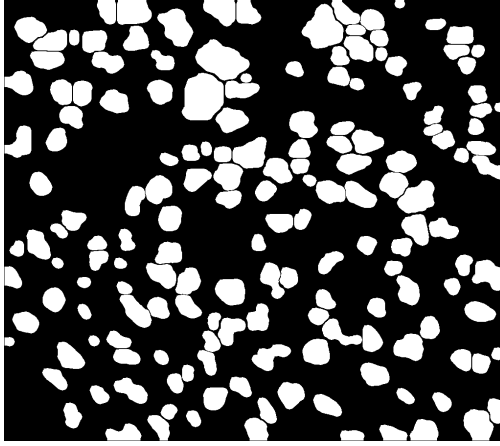
Step 3: Upgrade d_{next} as the maximum distance value of the pixels in the distance map that are not yet assigned a label of a regional marker.

Step 4: If $d_{next} \neq d_{min}$ then steps 2 and 3 are repeated.

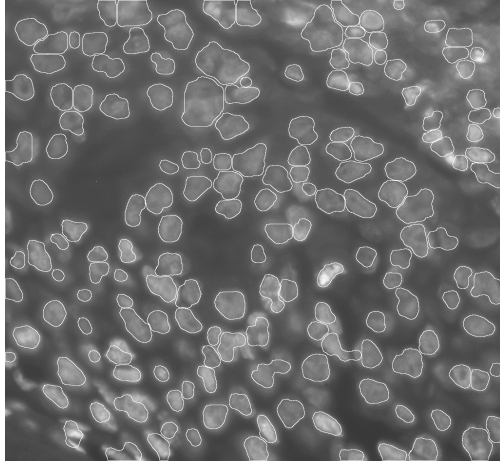
Figure 6 shows the result of multi-stage segmentation. Figure 6(a) is the two-tone version of segmented regions while Figure 6(b) shows overlaying of the boundary of isolated regions on the original image.

The artifacts in the segmented image are identified by measuring the size and the shape of the objects. From image I_{seg} that contains isolated cell nuclei, we have calculated size threshold values. All the objects which fall below the size threshold limit are considered as fragments and merged to the nearest larger object. If the fragment is connected to more than one object, then it is merged with that object with which it shares larger common boundary. Objects that are above the maximum size limit are removed from further processing. If the objects can be generally categorized as convex shapes, shape-filters can be incorporated to identify fragments. The standard deviation of the sampled boundary pixels from the object centroid is considered as a shape indicator. In our experiments, shape indicator did not alter the final result significantly. Shape based filters are effective in isolating non-nuclear artifacts or the cells whose shape change significantly due to malignancy. We argue

that the shape based filters should be a part of the integrated approach for making the segmentation process more global in nature.



(a)



(b)

Figure 6: Result of segmentation by region growing (a) Segmented image (b) Boundary of the segmented image is overlaid on the original grey scale image.

D. Shrink Wrapping

The accuracy of segmentation obtained by sequential combination of different region based techniques depends on the accuracy of thresholding and the intermediate processing stages. Generally, the shape of the segmented objects is influenced by the shape of the structuring elements used in the morphological operation for structural smoothing, noise reduction, etc. Segmentation of clusters relies on the presence of a concavity where two objects touch one another. These need not be the same pixel locations with local gradient maxima indicating the actual boundary separating touching cell-nuclei. Thus the shape of the cell nuclei obtained by

the above described region based processes might not depict the accurate shape of the cell-nuclei. To regain accurate shape, we implemented a dilation and shrink-wrap process. Here we consider the isolated objects extracted thus far as just the signatures of the cell nuclei. Each iteration step consists of growing signatures to occupy a pre-defined number of pixels in its immediate neighborhood followed by grey scale shrinking. The number of iterations of dilation-and-shrink defines the neighborhood region the signatures are grown into. Signatures are uniquely labeled prior to dilation to avoid merging of segmented regions. Grey scale shrinking is described as follows.

In the first step, the labeled nuclei-signatures are dilated by grey-scale morphological dilation.

In the second step, the nuclei-signatures are shrunk along their surface only if the surface/boundary pixels have a grey-level below a predefined threshold. Low grey scale spots entirely within the objects are not affected while the object shrinks.

During shrinking, all the pixels which are below certain threshold in its grey-level are converted into background under the condition that it does not create a hole within the nucleus-signature.

The process of shrinking continues until there are no more shrinking is possible i.e. all the boundary pixels have grey-level above the threshold. We have used the threshold $\varepsilon \cdot \mu_i$ where ε is a tuning constant and μ_i is the average intensity of the pixels that belong to the signature of the object i .

In the third step, the objects are dilated again and the process of shrinking is repeated. Ideally, the shrink-wrapping process is stopped when the difference between the size of the objects before and after an iteration of shrink-wrapping becomes zero. In practice, the process of shrink-wrapping is stopped when the change in size of the object between two consecutive iterations is negligible. Figure 7 shows the result of shrink-wrapping for improving the accuracy of segmentation.

If the objects are bloated beyond their actual boundary in the beginning of the shrink-wrap process, such objects are simply

shrunk back to their shape and no major change in the shape and size takes place after that. The advantage of this shrink-wrapping process is that the shrinking takes place based on the intensity characteristics of the individual cell nuclei. Thus in malignant tissues, individual nuclei that show different histogram features are not unduly influenced by the property of the pixels belonging to other cells.

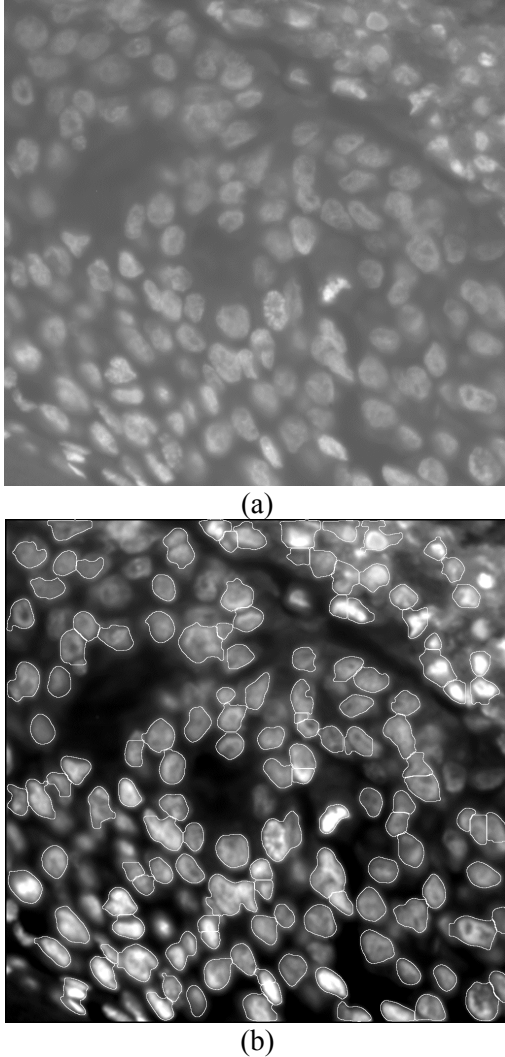


Figure 7: Result of segmentation (a) Original grey scale image, (b) After segmentation and shrink-wrapping.

The shrink-wrap process differs from region-growing in a significant way. Region growing process assumes that the initial seed or marker is well within the actual object boundary and hence does not allow shrinking. The shrink-wrap process allows the initial bloated region to shrink back to its shape. If the signature is not bloated uniformly around the object, shrink-wrap

process allows different parts of the boundary to either shrink or grow in the same iteration and facilitate accurate shape extraction.

III. Analysis of Genetic-Heterogeneity

Quantification of genetic variation in cancer is helpful to understand the genetic heterogeneity of the disease, which in turn can provide information about its history and more importantly, its future progression. Evidence of the existence of amplified and/or deleted genome regions has been extensively reported both in invasive carcinoma of the breast and DCIS [21]. To reduce the error due to loss of information by 2-D imaging, we have studied amplified genome regions instead of deleted ones as a marker of cancerous cells. The level of magnification is considered as a way to determine genetic heterogeneity between different parts of the tumor. The problem of loss of information due to 2-D imaging still exists for amplified signals, but it is much less significant compared to deleted signals.

In situ hybridization enables the morphological demonstration of specific genome or RNA sequences in individual cells, tissue sections or chromosome preparations. In our experiments, the fluorescent stained tissue sections are also stained by fluorescence *in situ* hybridization (FISH) with probes to the centromere and an area of chromosome 17 known to host the *Her2* gene, which is often over-expressed in DCIS. The *Her2* gene is labeled with a Red fluorescent tag while the centromeric probe was labeled green-labeled using a FITC fluorochrome. We propose to distinguish cancerous from normal cells using the number of copies of the *Her2* gene that are shown as bright spots in the FISH signal image. The centromere shows no amplification in malignancy and is used as a control to study the *Her2* amplification. One could argue that more accurate genetic analysis can be done using confocal imaging and 3-D analysis [22 23]. However, light scattering and hybridization efficiency limits the thickness of the sections that can be analyzed using confocal microscopy to less than 40 μ m. This is very small compared the usual thickness of our samples (3-5 mm).

This thickness is required to be able to preserve meaningful connected tissue structures such as ducts, etc. In addition, even if the light penetration and sample preparation allowed staining and imaging of such thick blocks, the amount of data produced would be multiple times higher than using physical sectioning and 2D imaging using a standard fluorescent microscope. Since we acquire information about the centromere of a chromosome that is not amplified due to malignancy, comparison of its quantified measure with the measure of the Her2 gene over large number cell-nuclei in the tissue can be expected to provide reliable information about genetic heterogeneity without the necessity of 3-D imaging. By analyzing cell nuclei of similar size and shape, we can argue that the error due to analyzing 2-D projections of what is essentially a 3-D structure, which is randomly located in a 3-D space within the cell nucleus, is considerably reduced. In summary, our argument favoring a standard 2D fluorescent microscope stems from factors such as the amount of data to be analyzed, complexity of the data and the reduced necessity of multi-dimensional imaging due to analysis of large number of cell nuclei.

To identify genetically aberrant cells one might count the number of fluorescent signals present in the cell nuclei or integrated the fluorescence intensity in the nucleus area. A reasonable method to detect FISH signals and to determine their parameters should be translation, scaling and rotation-invariant, and should be able to measure a range of parameters about the signal. The accuracy by which the parameters are determined must be as accurate as the level of noise permits. A simple algorithm for detecting the signals, which satisfies the above-mentioned conditions, is to locally threshold the image at an appropriate level and to characterize the signal by using its intensity, size and shape property to distinguish it from noise. A global top-hat filter is used to enhance 'spot'-like structures in the image prior to local thresholding and identifying FISH signals. The global top-hat also reduces the bleed-through effect in the different spectral channels of the image data.

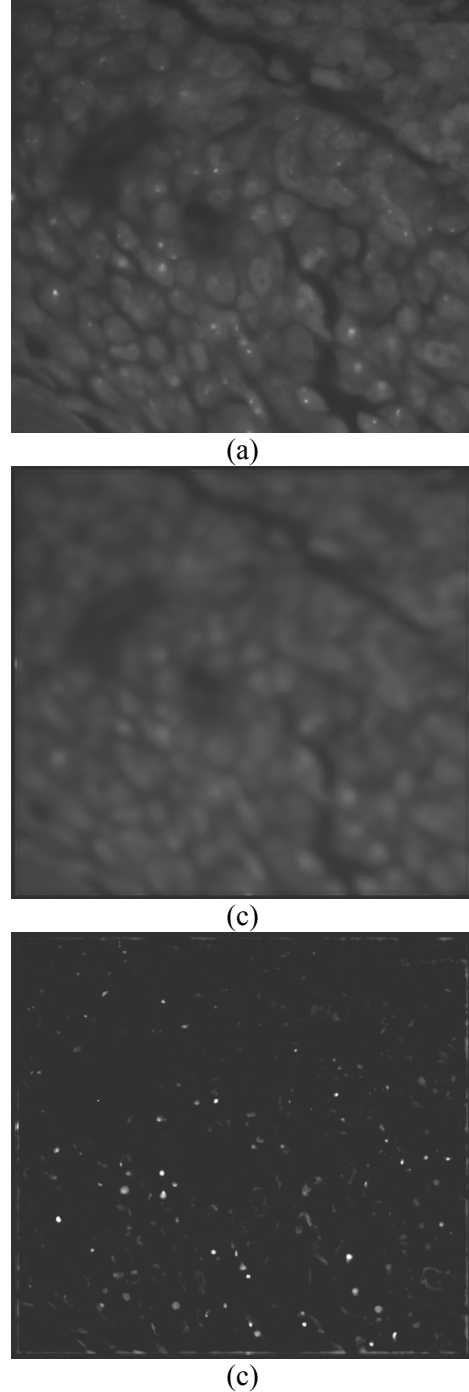


Figure 8: Result of processing FISH signal channel (a) Original image of the FISH signal channel (Her2 gene). The bright spots are the FISH signals (b) The background image representing bleed-through (c) After bleed-through reduction.

The region of interest for counting the FISH signals is only within the cell nuclei. For this purpose the segmented and labeled tissue image is virtually superimposed on each FISH signal channel. Regions outside the cell nuclei and regions of the truncated cell nuclei are discarded. All

the spot like structures in each FISH signal channel that correspond to nuclei regions are examined as to whether they are FISH signals or artifacts. Automatic detection of the overlapped signals is done by inspecting the intensity profiles and feature vector of the connected group of pixels. Analysis of overlapped Fluorescent signals is reported by Adiga *et al* [23].

After separating all the overlapped signals, the FISH signals are labeled separately for each cell nucleus. Figure 8, shows the result of processing of the FISH signal channel at successive stages. Figure 8(a) shows FISH spots representing *Her2* gene. Figure 8(b) is the representation of bleed-through calculated by global top-hat operation. Figure 8(c) shows the FISH signal channel after noise reduction. The number of signals present, average and relative brightness of the signals, average and relative size of the signals (in pixels), total brightness of the FISH signals, ratio of total size of the FISH signals in the nucleus to the size of the nucleus and the total fluorescence in each cell nuclei within each FISH channel are recorded for analyzing genetic heterogeneity.

IV. Results and Discussion

Experiments were conducted on a large number of data sets of thin tissue sections ($5\mu m$) of human breast cancer tissue. The sections were imaged using a Zeiss Axioplan I microscope. Image sizes varied depending on the size of the specimen and the areas being analyzed. To evaluate the segmentation we used one hundred and forty large tissue sections from two different cancer specimens. Our target, that 75% of the total cell nuclei be correctly segmented with less than 10% false positives, is achieved in almost all the cases.

The Importance of the accuracy in delineating nuclear boundary depends on the application. In cases such as cell counting, it is not really necessary to have an accurate marking of boundary or shape etc., as long as one would count the same number of cell-nuclei and accurately recognize an individual nucleus. It is important to have accurate shape information when we classify cells of different tissues or to distinguish healthy cell nuclei from a malignant ones, etc.

Segmentation accuracy can be measured as a percentage of symmetric difference in the area of the cell-nuclei given by automatic segmentation and manual segmentation. If ' A_1 ' is the area of a cell nucleus segmented by an automatic method, i.e. set of all pixels in the nucleus segmented by automatic method and ' A_2 ' is the area of the same nucleus segmented manually then the percentage of symmetric difference is given by
$$\frac{\{A_1 \cup A_2\} \setminus \{A_1 \cap A_2\}}{\# A_2} \times 100.$$
 Where

' \cup ' is the set union operator, ' \cap ' is the set intersection operator and ' \setminus ' is the set difference operator. One hundred cell-nuclei were randomly selected from different images and their boundaries manually delineated. The percentage symmetric difference in the area of each of the cell nuclei when segmented manually and automatically was then calculated. We have found that such a percentage difference never exceeded 7%, which is an acceptable error limit. Table 1 shows the total number of cells present, number of cells segmented, number of false positives in a few randomly selected data sets from three cases. The percentage scores shown in the table is a representative score and in all the images we have analyzed, we are mostly within the limits of at least 75% of the cell-nuclei segmentation with less than 10% false positive rate.

There is no generally accepted segmentation technique for delineating cell nuclei and hence is difficult to compare the proposed methodology to other state-of-the-art techniques. What we hope to achieve is that the proposed integrated approach and its variants would form the general segmentation technique that would be successful in most of the tissue analysis experiments.

Quantification of the number of copies of gene labeled using FISH is a straight forward process. We have analyzed two individual cases consisting of sixty-five and sixty tissue sections respectively. Twenty smaller size images are also analyzed. Approximately five to six distinct areas in each section are selected for acquiring higher magnification images. Each area of interest, on an average, contains three-thousand cell nuclei. Table 2 shows the number of data sets, the number of cell

nuclei that has amplified FISH signals and the corresponding standard deviation of the FISH signals per nucleus within a section and standard deviation of total fluorescence brightness per nucleus within a section (Dev. Total FISH). The standard deviation across the specimens shows that the genetic heterogeneity exists farther from the pre-malignant tumor.

Case	Sp. No.	Total Nuc.	Seg.Nuclei (%ge correct seg.)	False Positive s (%)
1.	1	5731	5538 (88%)	438(8%)
	2	4733	4333 (87%)	211(5%)
	3	4338	4691 (95%)	431(10%)
	4	3609	3655 (92%)	309(9%)
2	1	4064	4166 (94%)	327(8%)
	2	3171	3301 (87%)	523(16%)
	3	4008	3941 (89%)	372(10%)
	4	2089	2289 (93%)	328(15%)
	5	2007	2109 (93%)	230(11%)
	6	2061	2101 (93%)	183(9%)
3	1	65	68 (93%)	7 (11%)
	2	46	49 (86%)	9 (18%)

Table 1: Lists a selected number of images, total number of cells present,, number of segmented cells and the number of false positives.

Most the functions in the software are written in IDL. The C language is used for low-level processing of images. The software is tested on a 512 Mb RAM, 1GHz speed, Win2000 PC. An image of approximately 5K x 5K size takes about fifteen minutes to complete the segmentation and analysis of fluorescent signatures of the genes. To conclude, this paper has presented an automatic integrated methodology for segmenting cell populations in large tissue sections. The combination of the several segmentation techniques and enhancing the delineation features of the nuclei is shown to be effective even when the cells are overlapping or touching one another. By using shrink-wrapping as a post-processing technique, the segmentation technique is made more accurate and less dependent on the accuracy of initial thresholding. This automated approach for segmentation has the added advantage of repeatability of the experiment and reproducibility of the result. If any error exists, they will always be

systematic, and it will influence control and experimental populations identically.

Spec. 1	No.Cells	Dev. in No.FISH Signals	Dev. Total FISH
1	1749	0.465	1.36
2	1931	0.658	0.93
3	1856	0.839	2.11
4	2101	0.925	2.07
Avg. Dev.=		0.721	1.617
Spec. 2	No.Cells	Dev. in No.FISH Signals	Dev. Total FISH
1	1760	1.241	3.161
2	1720	1.02	3.015
3	931	1.13	1.923
4	2611	3.38	2.607
5	1176	1.307	2.431
6	1118	1.366	3.783
Avg. Dev.=		1.574	2.82
Spec. 3	No.Cells	Dev. In No.FISH Signals	Dev. Total FISH
1	26	2.103	4.036
2	23	2.404	3.664
Avg. Dev. =		2.253	3.85

Table 2: Shows the number of data sets (Spec.), number of isolated regions of interest (ROI) and the number of cell nuclei segmented from those regions of interest and corresponding deviation in number of FISH signals per cell nucleus.

Acknowledgement: Suggestions and advice from Dr. D. Knowels, Dr. D. Sudar , Dr. R. Malladi and Mr. Kevin of LBNL and Dr. S. Song of UC Berkeley are acknowledged.

References

- [1]. R Holland, JL Connolly, R Gelman, M Mravunac, JH Hendriks, AL Verbeek, SJ Schnitt, B Silver, J Boyages, and JR Harris, "The presence of an extensive intraductal component following a limited excision correlates with prominent residual disease in the remainder of the breast", J. of Clinical Oncology, Vol. 8, pp. 113-118, 1990.
- [2]. MD Lagios, PR Westdahl, FR Margolin, and MR Rose, "Ductal Carcinoma in Situ", Cancer, Vol. 40, pp. 1309-1314, 1982.
- [3]. K. Preston Jr., PH. Bartels, "Automated image processing for cell nuclei and tissue",

- In: Progress in Medical Imaging (Newhouse eds.), Springer Verlag, pp. 1-121, 1988.
- [4]. Gil J, Wu H, Wang BY., "Image analysis and morphometry in the diagnosis of breast cancer", *Microscopy Research and Techniques*, Vol. 59(2) pp. 109-18, 2002
- [5]. A. Nedzved, S. Ablameyko, Pitas I., "Morphological Segmentation of Histology Cell Images", *International Conference on Pattern Recognition (ICPR'00)*, Volume 1, Barcelona, 2000.
- [6]. Schupp S, Elmoataz A, Herlin P, Bloyet D., "Mathematical morphologic segmentation dedicated to quantitative immunohistochemistry", *Analytical Quantitative Cytology and Histology*, Vol. 23(4), pp. 257-67, 2001.
- [7]. Wu HS, Barba J, Gil J., "A parametric fitting algorithm for segmentation of cell images", *IEEE Trans. Biomedical Engineering*, Vol. 45(3), pp. 400-7, 1998.
- [8]. Yang Q, Parvin B., "Harmonic cut and regularized centroid transform for localization of subcellular structures", *IEEE Trans. Biomedical Engineering*, Vol. 50(4), pp. 469-75, 2003.
- [9]. Zimmer C, Labruyere E, Meas-Yedid V, Guillen N, Olivo-Marin JC., "Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing", *IEEE Trans. Medical Imaging*, Vol. 21(10), pp.1212-21, 2002.
- [10]. PS Umesh Adiga, "On Quantitative Evaluation of 3-D Histo-pathological Images Obtained using Confocal Microscope", Ph.D. thesis, Indian Statistical Institute, Calcutta, India, 1999.
- [11]. R. Fernandez-Gonzalez, A. Jones, E. Garcia-Rodriguez, P. Y. Chen, A. Idica, S. J. Lockett, M. H. Barcellos-Hoff, C. Ortiz-de-Solorzano, "A System for Combined Three-Dimensional Morphological and Molecular Analysis of Thick Tissue Specimens", *Microscopy Research and Techniques*, Vol. 59(6), pp. 522-530, 2002
- [12]. N. Sochen, R. Kimmel, R. Malladi, "A general framework for low level vision", *IEEE Transactions on Image Processing*, special issue on PDEs and Geometry-Driven Diffusion in Image Processing and Analysis, Vol. 7, pp. 310—318, 1998.
- [13]. R. Malladi, and Ravve I., 2002. "Fast difference schemes for edge enhancing Beltrami flow", In: Heyden et al. (Ed.), *Computer Vision ECCV 2002*, Vol. LNCS 2350, pp. 343-357, 2002.
- [14]. PK Sahoo, Soltani S., Wong AKC, Chen Y., "A Survey of Thresholding Techniques", *Computer Graphics and Image Processing*, Vol. 41, pp. 233-260, 1988.
- [15]. Umesh Adiga, B. B. Chaudhuri, "Efficient methods for noise reduction and enhancement of confocal microscopy images", *MICRON*, Vol. 32, pp. 363-370, 2001
- [16]. JN Kapoor, Sahoo PK, Wong AKC., "A new method for grey-level picture thresholding using the entropy of the histogram", *Graphical Models and Image Processing*, Vol. 29, pp. 273-285, 1985.
- [17]. A. Rosenfeld, De la Torre P., "Histogram concavity analysis as an aid to threshold selection", *IEEE Trans. on System Man and Cybernetics*, SMC-13, pp. 231-235, 1983.
- [18]. J.C. Russ, *The Image processing Handbook*, CRC Press, London, 1995.
- [19]. PS. Umesh Adiga and B. B. Chaudhuri, "An efficient method based on watershed and rule based merging for segmentation of 3-D histo-pathological images", *J. Pattern Recognition*, Vol. 34/7, pp-1449-1458, 2001.
- [20]. G. Borgefors., "On digital distance transforms in three dimensions", *Computer Vision Graphics and Image Processing*, Vol. 64, pp. 368-376, 1996.
- [21]. J.J Isola, "Genetic aberrations detected by comparative genomic hybridization predict outcome in node negative breast", *Am. J. Pathology*, Vol. 147, pp. 905-911, 1995.
- [22]. K. Rodenacker, M. Aubele, P. Hutzler and PS. Umesh Adiga, "Groping for quantitative 3D image analysis: an Approach to Quantitative Evaluation of Fluorescence in situ Hybridization in thick tissue sections of prostate carcinoma", *J. Analytical Cellular Pathology*, Vol. 15, pp. 19-29, 1997.
- [23]. PS. Umesh Adiga, Sam JL Knight, and B. B. Chaudhuri, "Characterization and automatic counting of FISH signals in 3-D tissue images", *J. Image Analysis and Stereology*, Vol. 20, No. 1, pp. 41-52, 2001.